Genome Sequence of the Yeast Clavispora lusitaniae Type Strain CBS 6936

Pascal Durrens, Christophe C. Klopp, Nicolas Biteau, Valérie Fitton-Ouhabi, Karine Dementhon, Isabelle Accoceberry, David J. Sherman, Thierry Noël

1 The genome of the yeast *Clavispora lusitaniae* type strain CBS 6936

2

3 Pascal Durrens[#abc], Christophe Klopp[e], Nicolas Biteau[bd], Valérie Fitton-Ouhabi[bd],

4 Karine Dementhon[bd], Isabelle Accoceberry[bd], David J. Sherman[cab], Thierry Noël[bd]

5

6 Laboratoire Bordelais de Recherche en Informatique (LaBRI), UMR-CNRS 5800,

7 Talence, France[a]; Université de Bordeaux, Bordeaux, France[b]; INRIA joint project

8 team PLEIADE[c]; Laboratoire de Microbiologie Fondamentale et Pathogénicité UMR-

9 CNRS 5234, Bordeaux, France[d]; Genotoul Bioinfo, MIAT, INRA Toulouse, Castanet-

10 Tolosan, France[e]

11

12

13 Running Head: Genome of *C. lusitaniae* type strain CBS 6936

14

15 #Address correspondence to Pascal Durrens, pascal.durrens@labri.fr

16

Abstract

Clavispora lusitaniae, an environmental saprophytic yeast belonging to the CTG clade of *Candida*, can behave occasionally as an opportunistic pathogen in humans. We report the genome sequence of the type strain CBS 6936. Comparison with sequences of strain ATCC 42720 indicates conservation of chromosomal structure but significant nucleotide divergence.

Clavispora lusitaniae, teleomorph of *Candida lusitaniae*, is an ubiquist environmental ascomycetous yeast, with no known specific ecological niche. It can be isolated from different substrates, such as soils, waters, plants, and gastrointestinal tracts of many animals including birds, mammals and humans. In immunocompromised hosts, *C. lusitaniae* can be pathogenic and is responsible for about 1% of invasive candidiasis, particularly in pediatric and onco-haematology patients (1).

So far, two strains have had their genomes sequenced : ATCC 42720, isolated from the blood of a patient with myeloid leukemia (2), and MTCC 1001, a self-fertile strain isolated from citrus (3). We report here the genome sequencing and assembly of the *C. lusitaniae* type strain CBS 6936 (4), isolated from citrus peel juice. Genomic DNA was isolated from a 50 ml YPD yeast culture, after sphaeroplast osmotic lysis and glass rod purification of ethanol-precipitated nucleic acids. The DNA library was prepared from 1 µg according to the NEBNext DNA Library Prep Master Mix Set for Illumina (E6040) protocol with an insert size of 368±122nt generated with an ultrasonicator Covaris.

The library was sequenced on a MiSeq Illumina v1.18.54 producing *circa* 1.5 million paired end reads, representing a 28X coverage. A *de novo* assembly of the reads was performed with the aim to compare the contigs to those of the ATCC 42720 strain. We used SPAdes v3.9.0 with default parameters (5) to perform the assembly. Only the 53 scaffolds longer than 1kb were retained as the GC % of many shorter ones deviated from the average.

The scaffolds represent an overall length of 12.0Mb which is close to the 12.11Mb of the ATCC 42720 genome. The GC% amounts to 44.5% which is identical to that of ATCC 42720 strain. The alignment of CBS 6936 scaffolds against ATCC 42720 contigs by Nucmer (6) shows almost co-linearity between the two sets, apart from a 66kb long inversion. The number of SNPs between CBS 6936 and ATCC 42720 amounts to 132141 SNPs with Phred quality higher than 30, according to BWA (7) alignment of reads and Freebayes detection (8). This SNP density of 1 SNP per 90bp is twice the level observed between the strains SC5314 and WO-1 of *Candida albicans*, members of different subgroups in the species and qualified as relatively recent divergent (9).

Genes for 4 rRNA, 197 tRNA and 5539 proteins were predicted by alignment, TRNAscan-SE and a combination of Augustus (10), Snap (11) and Genmark-ES (12), respectively. The average identity of proteins between the two strains amounts to 89.9% ± 23.2, based on BLASTp (13) best hit alignments.

Altogether, these results indicate that, in spite of a conserved genome structure, the sequences are different, meaning that the strains underwent a significant divergence.

Accession numbers. This whole genome shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession numbers SRP075809 for reads and LYUB00000000 for scaffolds. The second version of the assembly of CBS 6936 is version LYUB02000000 and is described here.

Acknowledgements

The library and nucleotide sequencing were performed at the Genomic and Transcriptomic Facility of Bordeaux (grants from the Conseil Regional d'Aquitaine n°20030304002FA and 20040305003FA and from the European Union, FEDER n°2003227 and from Investissements d'avenir, Convention attributive d'aide N°ANR-10-EQPX-16-01). We would like to thank Christophe Hubert for its expertise and technical assistance.

The development of the bioinformatic pipeline was carried out using the PLAFRIM experimental testbed, being developed under the Inria PlaFRIM development action with support from Bordeaux INP, LABRI and IMB and other entities: Conseil Régional d'Aquitaine, Université de Bordeaux and CNRS (and ANR in accordance to the programme d'investissements d'Avenir (see https://plafrim.fr/).

References

1. **Favel A, Michel-Nguyen A, Peyron F, Martin C, Thomachot L, Datry A, Bouchara JP, Challier S, Noël T, Chastin C, Regli P**. 2003. Colony morphology switching of *Candida lusitaniae* and acquisition of multidrug resistance during treatment of a renal infection in a newborn: case report and review of the literature. Diagn. Microbiol. Infect. Dis. **47**:331-339.
2. **Pappagianis D, Collins MS, Hector R, Remington J.** 1979. Development of resistance to amphotericin B in *Candida lusitaniae* infecting a human. Antimicrob Agents Chemother. 16:123-126.
3. **Lachance M**, **Daniel H**, **Meyer W**, **Prasad G**, **Gautam S**, **Boundymills K**. 2003. The D1/D2 domain of the large-subunit rDNA of the yeast species is unusually polymorphic. FEMS Yeast Res **4**:253–258.
4. **Rodrigues de Miranda L.** 1979. *Clavispora,* a new yeast genus of Saccharomycetales. Antonie van Leeuwenhoek. **45**:479-483.
5. **Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA.** 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol. **5**:455-477.
6. **Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL.** 2004. Versatile and open software for comparing large genomes. Genome Biology, 5:R12.
7. **Li H, Durbin R.** 2009. Fast and accurate short read alignment with Burrows-

Wheeler Transform. Bioinformatics, 25:1754-1760.

8. **Garrison E, Marth G.** 2012. Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907 [q-bio.GN].

9. **Butler G, Rasmussen MD, Lin MF, , Santos MAS, Sakthikumar S, Munro CA, Rheinbay E, Grabherr M, Forche A, Reedy JL, Agrafioti I, Arnaud MB, Bates S, Brown AJP, Brunke S, Costanzo MC, Fitzpatrick DA, de Groot PWJ, Harris D, Hoyer LL, Hube B, Klis FM, Kodira C, Lennard N, Logue ME, Martin R, Neiman AM, Nikolaou E, Quail MA, Quinn J, Santos MC, Schmitzberger FF, Sherlock G, Shah P, Silverstein K, Skrzypek MS, Soll D, Staggs R, Stansfield I, Stumpf MPH, Sudbery PE, Thyagarajan S, Zeng Q, Berman J, Berriman M, Heitman J, Gow NAR, Lorenz MC, Birren BW, Kellis M, Cuomo CA.** 2009. Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. Nature 459:657-662.

10. **Stanke M, Schöffmann O, Morgenstern B, Waack S.** 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics 7, 62.

11. **Korf I.** 2004. Gene finding in novel genomes. BMC Bioinformatics 5, 59.

12. **Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M.** 2008. Gene prediction in novel fungal genomes using an *ab initio* algorithm with unsupervised training. Genome Res. 18:1979-1990.

13. **Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL.** 2008 BLAST+: architecture and applications. BMC Bioinformatics 10:421.